Formation Data Scientist OpenClassrooms

Projet 4 – Livrable 3 – Support de présentation

Anticipez les besoins en consommation électrique de bâtiments

Etudiant : Monine Chan
Evaluateur : Late Lawson

Mardi 09 Novembre 2021
Version 2

PLAN

1. Présentation de la problématique, interprétation et pistes de recherche

2. Nettoyage des données et exploration.

3. Modélisations effectuées

4. Présentation du modèle final sélectionné

PLAN

1. Présentation de la problématique, interprétation et pistes de recherche

2. Nettoyage des données et exploration.

3. Modélisations effectuées

4. Présentation du modèle final sélectionné

1. PRÉSENTATION DE LA PROBLEMATIQUE

CONTEXTE

- Nous avons accès à des jeux de données qui contiennent des mesures de consommation d'énergie et d'émissions de CO2 pour certains bâtiments de la ville de Seattle en 2015 et 2016.

- Le but de ce projet est d'utiliser ces jeux de données pour prédire :
  - la consommation d'énergie,
  - les émissions de CO2

pour les bâtiments qui n'ont pas encore été mesurés.

- Nous devons également évaluer si l'ENERGY STAR Score est utile pour prédire les émissions de CO2.

1. PRÉSENTATION DE LA PROBLEMATIQUE

INTERPRETATION ET PISTES DE RECHERCHE

- Après avoir nettoyé le jeu de données, nous allons sélectionner les variables qui sont pertinentes pour calculer l'énergie et les émissions de CO2.

- Nous allons regarder s'il existe de grosses différences entre 2015 et 2016 en moyenne pour l'énergie et les émissions de CO2. Si non, on créera un jeu de données qui représente la moyenne sur les deux années.

- On fera ensuite une analyse univariée pour se familiariser avec le jeu de données.

- Enfin, on utilisera ce jeux de données pour tester différents modèles qui permettront de prédire:
  - la consommation d'énergie,
  - les émissions de CO2 sans l'ENERGY STAR Score. La raison est que l'ENERGY STAR Score est par définition calculé sur la consommation d'énergie. Or la consommation d'énergie est elle-même corrélée à un certain degré avec les émissions de CO2 donc il y a un risque de fuite de données si une feature (variable) est elle-même dépendante d'autre features (variables).

PLAN

1. Présentation de la problématique, interprétation et pistes de recherche

2. Nettoyage des données et exploration.

3. Modélisations effectuées

4. Présentation du modèle final sélectionné

2. NETTOYAGE DES DONNÉES

VARIABLES (COLONNES) POUR MODÉLISER L'ENERGIE

- **Qualitatives (x2) :**
  - ✓ ***PrimaryPropertyType*** : Type d'usage d'une propriété (bureaux, magasin, etc.)
  - ✓ ***LargestPropertyUseType*** : Type d'usage de la surface la plus importante d'une propriété (bureaux, magasin, etc).

- **Quantitatives (x7):**
  - ✓ ***NumberofBuildings*** : Nombre de bâtiments d'une propriété.
  - ✓ ***NumberofFloors*** : Nombre d'étages.
  - ✓ ***PropertyGFATotal*** : Surface totale du bâtiment et des parkings.
  - ✓ ***YearBuilt*** : Année de construction.
  - ✓ ***LargestPropertyUseGFA*** : Surface de la partie la plus utilisée du bâtiment.
  - ✓ ***SecondLargestPropertyUseGFA*** : Surface de la 2ème partie la plus utilisée du bâtiment.
  - ✓ ***ThirdLargestPropertyUseGFA*** : Surface de la 3ème partie la plus utilisée du bâtiment.

2. NETTOYAGE DES DONNÉES

VARIABLES (COLONNES) POUR MODÉLISER LES EMISSIONS DE CO2

- **Qualitatives (x2) :**
  - ✓ ***PrimaryPropertyType*** : Type d'usage d'une propriété (bureaux, magasin, etc.)
  - ✓ ***LargestPropertyUseType*** : Type d'usage de la surface la plus importante d'une propriété (bureaux, magasin, etc).

- **Quantitatives (x6):**
  - ✓ ***NumberofBuildings*** : Nombre de bâtiments d'une propriété.
  - ✓ ***NumberofFloors*** : Nombre d'étages.
  - ✓ ***PropertyGFABuilding*** : Surface totale en pieds carrés entre les murs extérieurs d'un bâtiment.
  - ✓ ***PropertyGFAParking*** : Surface totale en pieds carrés des parkings (fermés, partiellement fermés, ouverts).
  - ✓ ***YearBuilt*** : Année de construction.
  - ✓ ***LargestPropertyUseGFA*** : Surface de la partie la plus utilisée du bâtiment.

2. NETTOYAGE DES DONNÉES

SELECTION DES INDIVIDUS (LIGNES)

- Détection des valeurs aberrantes :
  - ✓ Les surfaces sont des nombres positifs donc on supprime toutes les lignes pour lesquelles les valeurs de PropertyGFABuilding(s) et PropertyGFAParking sont strictement négatives.

- Détection des lignes en doublons:
  - ✓ On vérifie que l'on n'a pas de doublons grâce à une fonction qui va compter les lignes en doublon.

```
Entrée [163]: dataset_overview(df_joined_cleaned)

Nombre de colonnes : 25
Nombre de lignes : 1669
Nombre de NaN : 2124
Pourcentage de NaN (%) : 5.09%
Nombre de colonnes en doublon : 0
Nombre de lignes en doublon : 0
Pourcentage de lignes en doublon (%) : 0.00%
```

2. NETTOYAGE DES DONNÉES

FEATURE ENGINEERING

- Détection des valeurs aberrantes :
  - ✓ Les surfaces sont des nombres positifs donc on supprime toutes les lignes pour lesquelles les valeurs de PropertyGFABuilding(s) et PropertyGFAParking sont strictement négatives.

- Différence entre le jeu de données 2015 et 2016:
  - On réalise un group by par identifiant des bâtiments (***TaxParcelIdentificationNumber***) et on effectue une moyenne entre 2015 et 2016.

  - On constate que l'énergie consommée et les émissions de CO2 augmentent de +8% environ entre 2015 et 2016 mais ceci est dû à l'apparition de 36 nouveaux bâtiment mesurés qui rajoute +6.4% en surface totale.

2. NETTOYAGE DES DONNÉES

RÉSUMÉ

➢ Comparaison avant et après nettoyage & sélection :

Proprieté du jeu de données	Avant nettoyage et sélection	Après nettoyage et sélection
Nombre de variables (colonnes)	56	24
Nombre d'individus (lignes)	6716	3668
Nombre d'éléments (colonnes x lignes)	312 276	88 032
Nombre de NaN	46 464	17 440
Pourcentage de NaN	14.8%	19.81%

NB: On a gardé dans le jeu de données nettoyées plus de colonnes que celles qui ont servi à faire les modèles. En effet, il peut arriver que la recherche de la modélisation impose de modifier les variables d'entrées pour améliorer la performance du modèle.

2. EXPLORATION

ANALYSE UNIVARIÉE : TENDANCE CENTRALE ET DISPERSION

- Mesures de tendance centrale et dispersion pour les variables qualitatives en entrée des modèles.

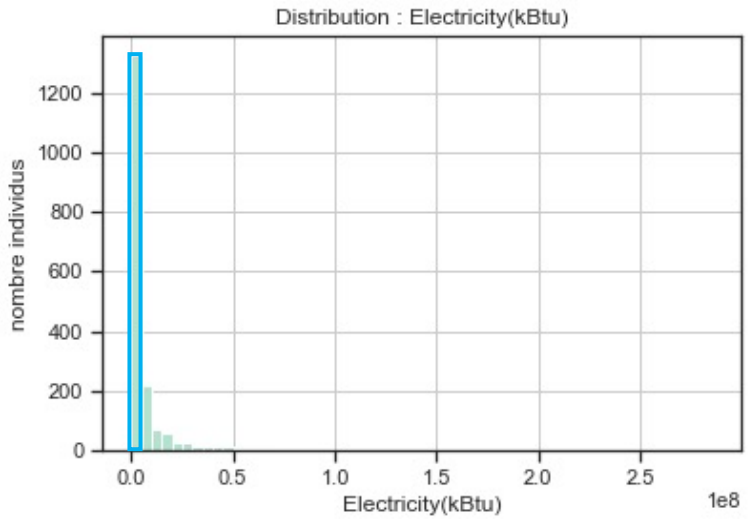
	Electricity (kBtu)	LargestPropertyUseTypeGFA	NumberofBuildings	NumberofFloors	PropertyGFABuilding(s)	PropertyGFAParking	PropertyGFATotal	SecondLargestPropertyUseTypeGFA	SiteEnergyUse(kBtu)	ThirdLargestPropertyUseTypeGFA	TotalGHGEmissions	YearBuilt
Moyenne	6,02E+06	1,01E+05	1,085938	4,704739	1,08E+05	15357,03344	1,23E+05	39404,47741	**8,50E+06**	14710,81418	**183,269387**	1960,8
Ecart-type	1,49E+07	1,70E+05	0,912837	7,501207	1,81E+05	45469,34683	2,07E+05	70293,11959	**2,19E+07**	34852,38035	**657,35506**	32,8
min	0,00E+00	6,46E+03	0	0	1,09E+04	0	2,00E+04	0	**0,00E+00**	0	**0**	1900
1er Quartile	7,54E+05	2,59E+04	1	1	2,91E+04	0	3,02E+04	6149,75	**1,26E+06**	2827,75	**20,465**	1929
Médiane	1,71E+06	4,52E+04	1	3	4,93E+04	0	5,13E+04	13124	**2,61E+06**	6129	**53,8475**	1964,6
3ème Quartile	5,27E+06	9,75E+04	1	5	1,02E+05	0	1,16E+05	36131,25	**7,46E+06**	13364,25	**154,10875**	1988
Max	2,85E+08	1,72E+06	20	99	2,20E+06	512608	2,20E+06	686750	**4,48E+08**	459748	**16870,98**	2015

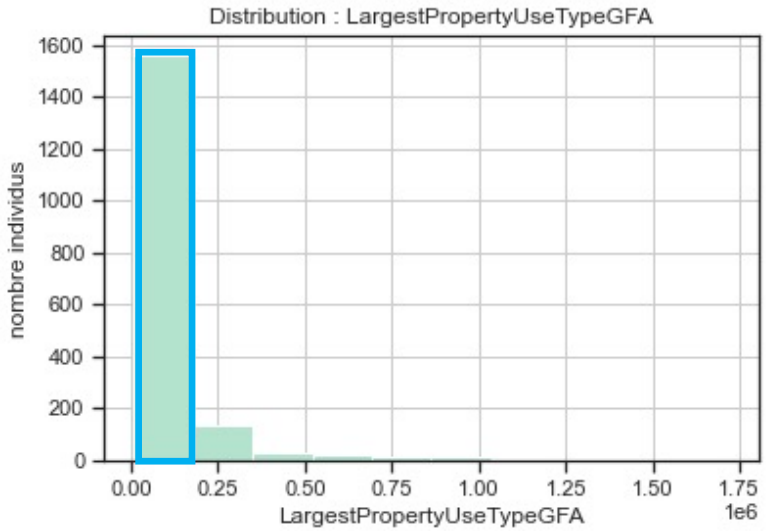
Targets à prédire

2. EXPLORATION

ANALYSE UNIVARIÉE : DISTRIBUTION DES VARIABLES

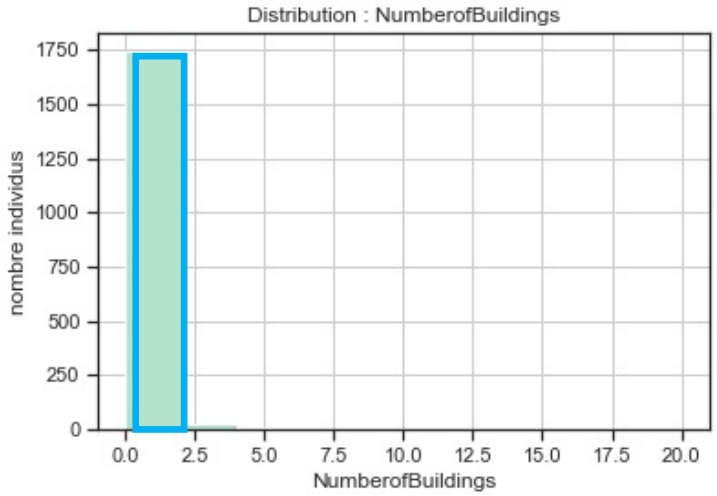
mode = 0
skew = 9.4 > 0
étalée à droite
kurtosis = 140 > 0
concentrée

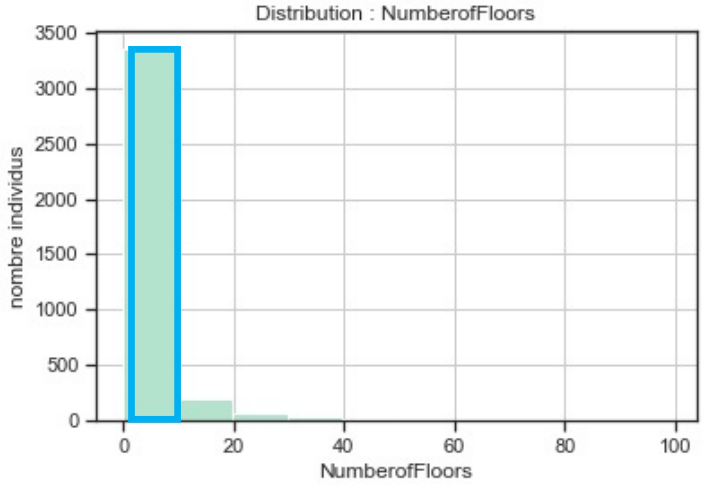




mode = 21600
skew = 4,64 > 0
étalée à droite
kurtosis = 28,2 > 0
concentrée

mode = 1
skew = 13.3 > 0
étalée à droite
kurtosis = 209 > 0
concentrée



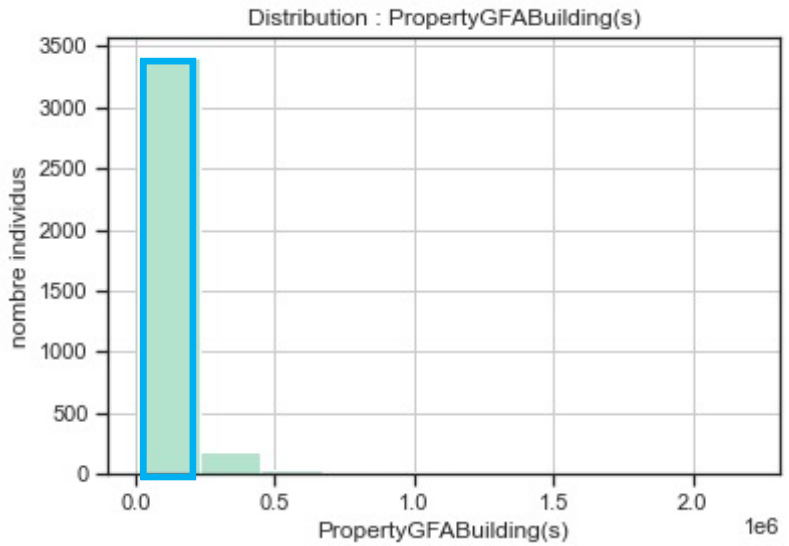


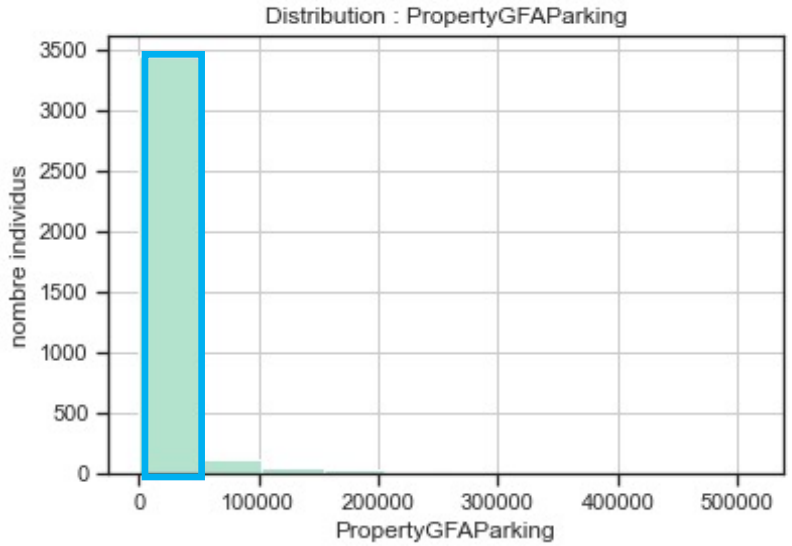
mode = 3
skew = 5.4 > 0
étalée à droite
kurtosis = 44 > 0
aplatie

2. EXPLORATION

ANALYSE UNIVARIÉE : DISTRIBUTION DES VARIABLES

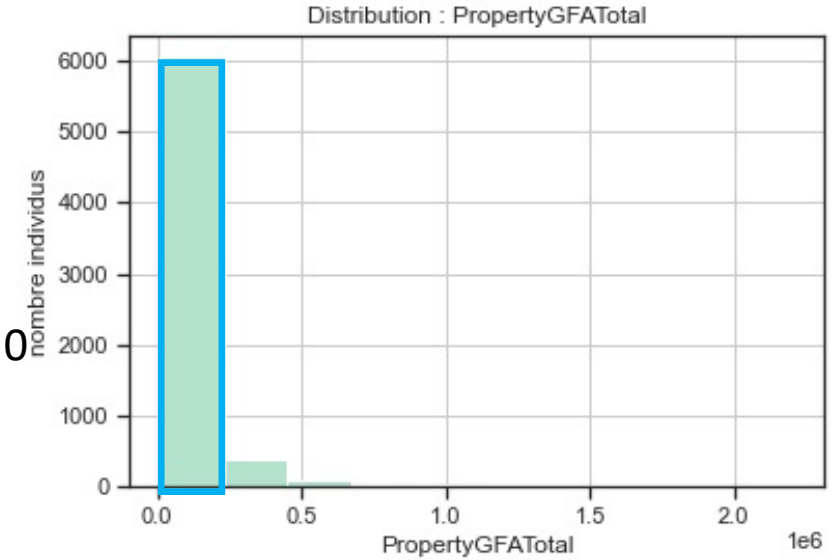
mode = 21600
skew = 5.96 > 0
étalée à droite
kurtosis = 51.4 > 0
concentrée

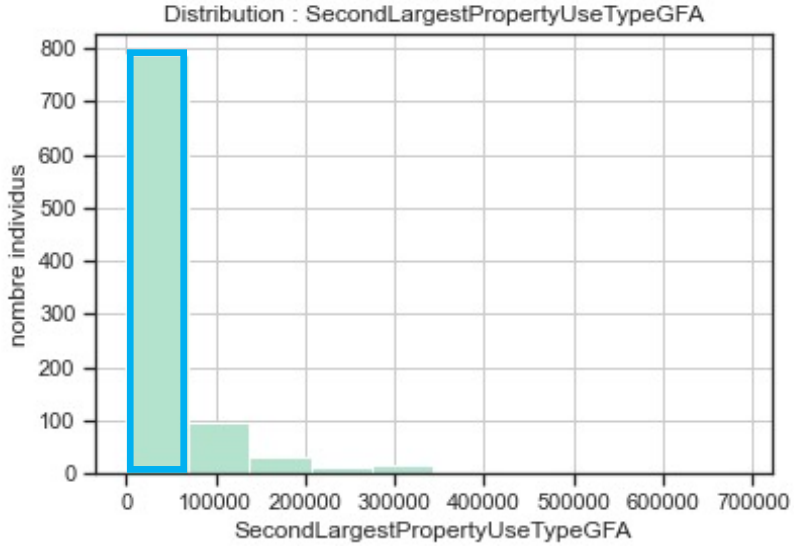




mode = 0
skew = 5.83 > 0
étalée à droite
kurtosis = 45.8 > 0
concentrée

mode = 21600
skew = 5.52 > 0
étalée à droite
kurtosis = 42.8 > 0
concentrée



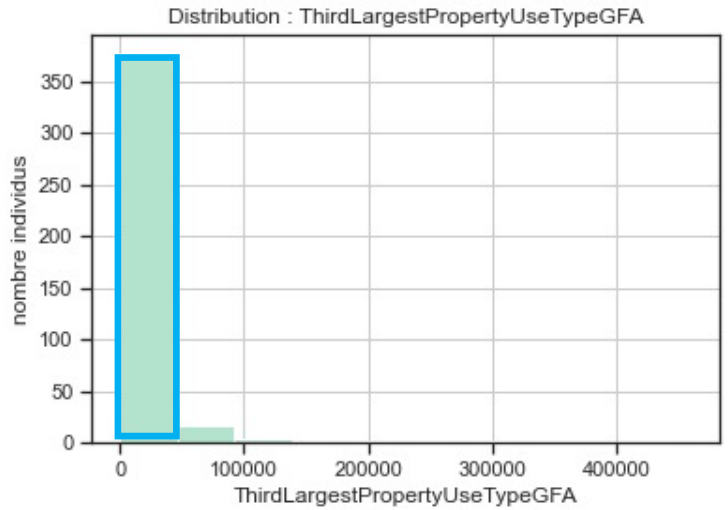


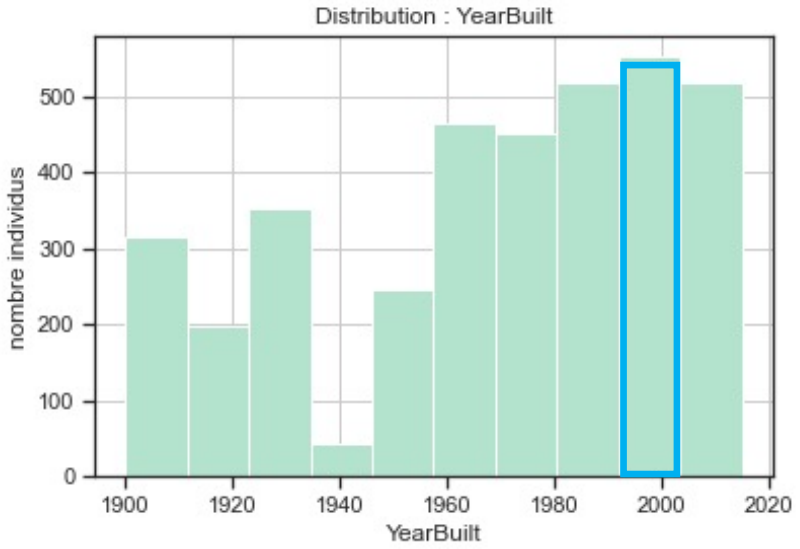
mode = 0
skew = 4,68 > 0
étalée à droite
kurtosis = 31,9 > 0
concentrée

2. EXPLORATION

ANALYSE UNIVARIÉE : DISTRIBUTION DES VARIABLES

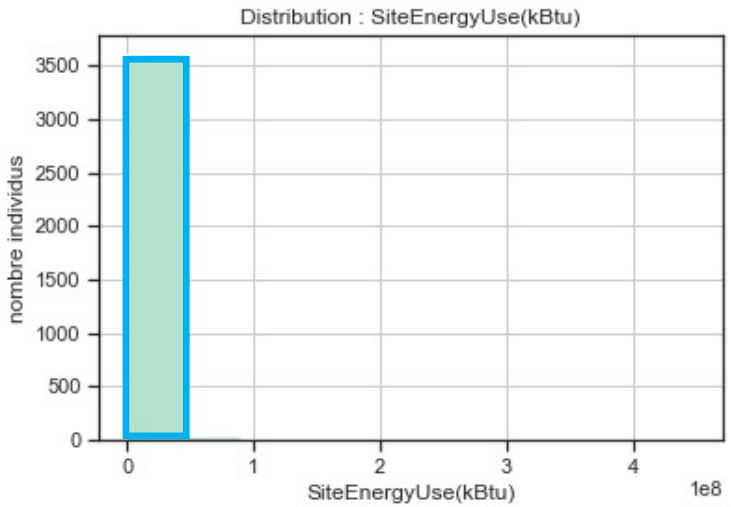
mode = 0
skew = 9.42 > 0
étalée à droite
kurtosis = 123 > 0
concentrée

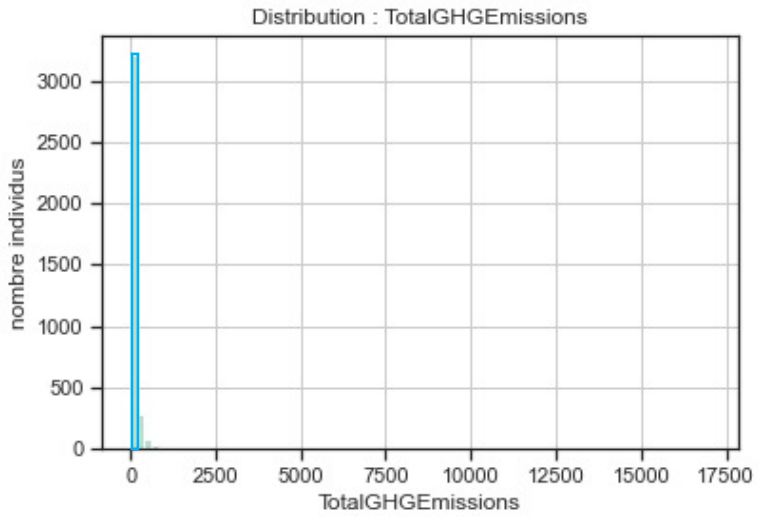




mode = 2000
skew = -0,52 < 0
étalée à gauche
kurtosis = -0.93 < 0
aplatie

mode = 0
skew = 13.65 > 0
étalée à droite
kurtosis = 275 > 0
concentrée

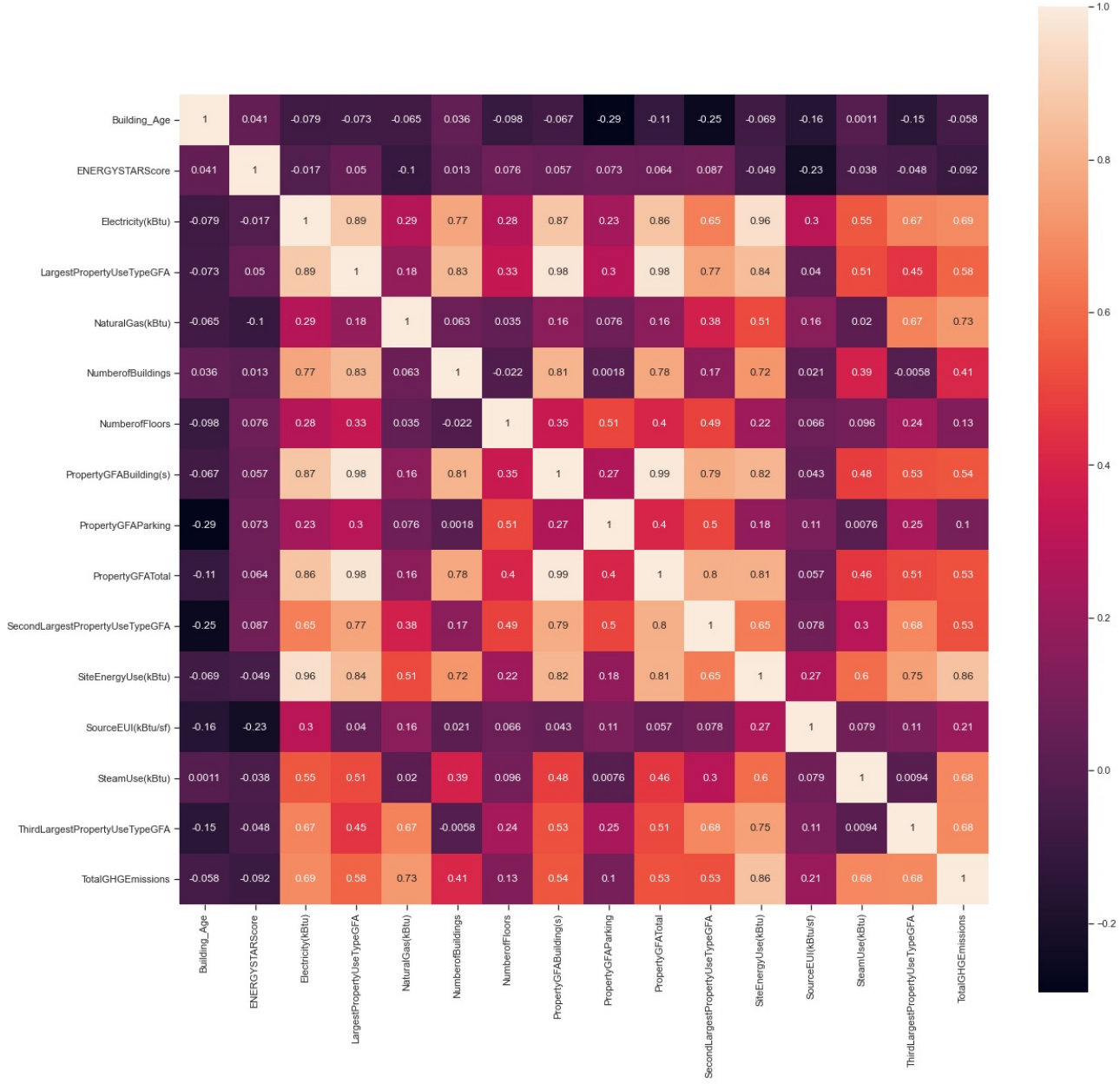




mode = 0
skew = 15.3 > 0
étalée à droite
kurtosis = 315
concentrée

2. EXPLORATION

ANALYSE BIVARIÉE : CORRÉLATIONS ENTRE VARIABLES QUALITATIVES



- On constate que **l'énergie consommée ('SiteEnergyUse(kBtu)')** a une forte corrélation avec les variables suivantes :
  - R^2 = 0.96 : l'électricité consommée : Electricity(kBtu)
  - R^2 = 0.86 : Les émissions de CO2 : TotalGHGEmissions
  - R^2 = 0.84 : la surface de 1ère utilisation du bâtiment : LargestPropertyUseTypeGFA
  - R^2 = 0.82 : la surface du bâtiment : PropertyGFABuilding(s)
  - R^2 = 0.81 : la surface totale (bâtiment et parking) : PropertyGFATotal
  - R^2 = 0.75 : la surface de 3ème utilisation du bâtiment : ThirdLargestPropertyUseTypeGFA
  - R^2 = 0.72 : le nombre de bâtiments : Number of Buildings. Attention, la majorité des individus n'ont qu'un seul bâtiment.
  - R^2 = 0.65 : la surface de 2ème utilisation du bâtiment : SecondLargestPropertyUseTypeGFA
- On constate que **les émissions de CO2 ('TotalGHGEmissions')** a une forte corrélation avec les variables suivantes :
  - R^2 = 0.73 : le gaz naturel consommé : NaturalGas(kBtu)*
  - R^2 = 0.69 : l'électricité consommée : Electricity(kBtu)*
  - R^2 = 0.68 : la vapeur consommée : SteamUse(kBtu)*
  - R^2 = 0.68 : la surface de 3ème utilisation du bâtiment : ThirdLargestPropertyUseTypeGFA
  - R^2 = 0.58 : la surface de 1ère utilisation du bâtiment : LargestPropertyUseTypeGFA
  - R^2 = 0.54 : la surface du bâtiment : PropertyGFABuilding(s)
  - R^2 = 0.53 : la surface de 2ème utilisation du bâtiment : SecondLargestPropertyUseTypeGFA
  - R^2 = 0.53 : la surface totale (bâtiment et parking) : PropertyGFATotal
  - *= déjà comptée dans l'énergie consommée : SiteEnergyUse(kBtu))*
- **L'ENERGY STAR Score est peu corrélé avec les émissions de CO2** (idem pour la consommation d'énergie).

2. EXPLORATION

ANALYSE BIVARIÉE : CONCLUSIONS

- Les énergies consommées et les émissions de CO2 sont fortement correlées avec les **surfaces** des bâtiments.

- Les différents types d'énergie (électricité, vapeur, gaz naturel) sont fortement corrélés à l'énergie consommée mais cela n'est pas surprenant puisque leur somme est égale à l'énergie consommée (totale).

- Les émissions de CO2 sont fortement corrélées aux consommations d'énergies.

PLAN

1. Présentation de la problématique, interprétation et pistes de recherche

2. Nettoyage des données et exploration.

3. Modélisations effectuées

4. Présentation du modèle final sélectionné

3. MODELISATIONS EFFECTUEES

PRINCIPES DES MODELISATIONS RÉALISÉES

- On crée un jeu d'entraînement(70%) et un jeu de test (30%). Les variables d'entrées qualitatives et quantitatives (X) sont définies :
  - au slide 7 pour l'énergie et
  - respectivement au slide 8 pour les émissions de CO2.
- Les targets (y) sont :
  - l'énergie « SiteEnergyUse(kBtu) » et
  - respectivement les émissions de CO2 « TotalGHGEmissions ».

- On effectue une stratification des données selon la variable « PrimaryPropertyType ».

- La performance des modèles de machine learning sont définies par deux critères :
  - RMSE : Root Mean Squared Error
  - R^2 : coefficient de détermination.

3. MODELISATIONS EFFECTUEES

PRINCIPES DES MODÉLISATIONS RÉALISÉES

- On effectue une transformation des variables d'entrées :
  - qualitatives : on effectue un encodage avec One Hot Encoder
  - quantitatives : on normalise les valeurs avec StandardScaler avec un SimpleImputer qui remplace les NaN par la médiane.
  - Ces opérations sont effectuées via un pipe-line.

- On teste ensuite les modèles de Machine Learning suivants et on sélectionne le plus performant parmi ces 2 familles de modèles:
  - Linéaire : Régression linéaire, Ridge, Lasso
  - Non linéaire : SVM.
  - Ensemblistes : Random Forest, XGBoost, GradientBoostingRegressor.

- Une fois le modèle de ML choisi, on effectue une recherche d'hyperparamètres pour optimiser le résultat:
  - On réalise une optimisation sur une grille de recherche couplée à une validation croisée.
  - On fait ensuite une prédiciton sur le jeu de test pour vérifier la cohérence de la performance du modèle optimisé avec ses hyperparamètres.

3. MODÉLISATIONS EFFECTUÉES

3.1. MODÉLISATION DE L'ÉNERGIE CONSOMMÉE

➢ On teste ici plusieurs modèles et on se basera sur la RMSE la plus faible et le R^2 le plus proche de 1 pour choisir le modèle.

	Dummy Regressor	Linear Regression	Ridge	Lasso	SVM	Random Forest	XGBoost	GradientBoostingRegressor
RMSE	2.003684e+07	1.003649e+19	1.343012e+07	1.310292e+07	2.003681e+07	1.638028e+07	2.184854e+07	1.674608e+07
RMSE/mean	2.357081e+00	1.180667e+12	1.579884e+00	1.541393e+00	2.357078e+00	1.926933e+00	2.570205e+00	1.969965e+00
RMSE/median	7.669668e+00	3.841753e+12	5.140758e+00	5.015514e+00	7.669658e+00	6.270016e+00	8.363150e+00	6.410038e+00
R^2	-8.600000e-02	-2.726022e+23	5.120000e-01	5.350000e-01	-8.600000e-02	2.740000e-01	-2.920000e-01	2.410000e-01

C'est le Lasso qui est le meilleur modèle car il fournit:

- la RMSE la plus faible : $1.31e^{+07}$
- le coefficient de détermination $R^2 = 0.53$ le plus élevé.

3. MODÉLISATIONS EFFECTUÉES

3.2. MODÉLISATION DES ÉMISSIONS DE CO2 SANS ENERGYSTARSCORE

➢ On teste ici plusieurs modèles et on se basera sur la RMSE la plus faible et le R^2 le plus proche de 1 pour choisir le modèle.

	Dummy Regressor	Linear Regression	Ridge	Lasso	SVM	Random Forest	XGBoost	GradientBoostingRegressor
RMSE	485.451000	1.891808e+14	319.838000	311.869000	474.405000	221.953000	460.544000	199.378000
RMSE/mean	2.648840	1.032256e+12	1.745179	1.701699	2.588565	1.211075	2.512934	1.087897
RMSE/median	9.015298	3.513271e+12	5.939697	5.791715	8.810152	4.121880	8.552743	3.702644
R^2	-0.053000	-1.599137e+23	0.543000	0.565000	-0.006000	0.780000	0.052000	0.822000

C'est le GradientBoostingRegressor qui est le meilleur modèle car il fournit:

- la RMSE la plus faible : 199,4.
- le coefficient de détermination $R^2 = 0.82$ le plus élevé.

3. MODÉLISATIONS EFFECTUÉES

3.2. MODÉLISATION DES ÉMISSIONS DE CO2 AVEC ENERGYSTARSCORE

➢ On teste ici plusieurs modèles et on se basera sur la RMSE la plus faible et le R^2 le plus proche de 1 pour choisir le modèle.

	Dummy Regressor	Linear Regression	Ridge	Lasso	SVM	Random Forest	XGBoost	GradientBoostingRegressor
RMSE	485.451000	1.891808e+14	319.838000	311.869000	474.405000	221.953000	460.544000	198.746000
RMSE/mean	2.648840	1.032256e+12	1.745179	1.701699	2.588565	1.211075	2.512934	1.084449
RMSE/median	9.015298	3.513271e+12	5.939697	5.791715	8.810152	4.121880	8.552743	3.690911
R^2	-0.053000	-1.599137e+23	0.543000	0.565000	-0.006000	0.780000	0.052000	0.824000

C'est le GradientBoostingRegressor qui est le meilleur modèle car il fournit:

- la RMSE la plus faible : 198,7.
- le coefficient de détermination $R^2 = 0.82$ le plus élevé.

PLAN

1. Présentation de la problématique, interprétation et pistes de recherche

2. Nettoyage des données et exploration.

3. Modélisations effectuées

4. Présentation du modèle final sélectionné

4. MODÈLE FINAL SELECTIONNÉ

OPTIMISATION DES HYPERPARAMETRES

- Une fois le modèle de ML choisi, on effectue une recherche d'hyperparamètres pour optimiser le résultat:
  - On réalise une optimisation sur une grille de recherche couplée à une validation croisée.
  - On fait ensuite une prédiction sur le jeu de test pour vérifier la cohérence de la performance du modèle optimisé avec ses hyperparamètres.

4. MODÈLE FINAL SELECTIONNÉ POUR L'ÉNERGIE CONSOMMÉE

4.1. MODÈLE LASSO OPTIMISÉ

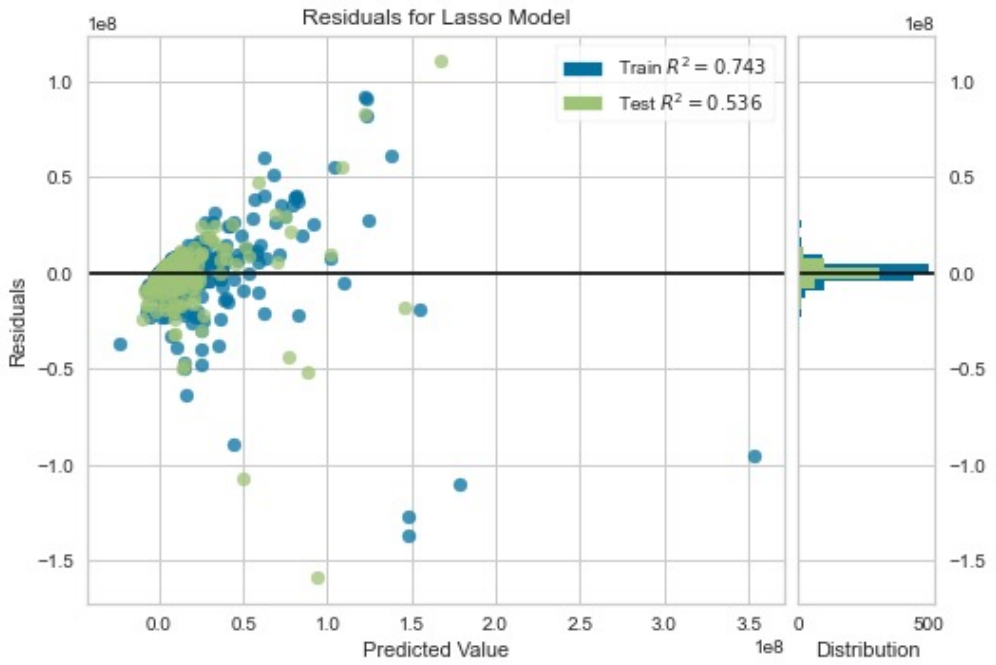
- On cherche l'hyperparamètre alpha/lambda du Lasso (paramètre qui limite l'overfitting du modèle) pour affiner la performance du modèle avec un scoring défini par la RMSE.

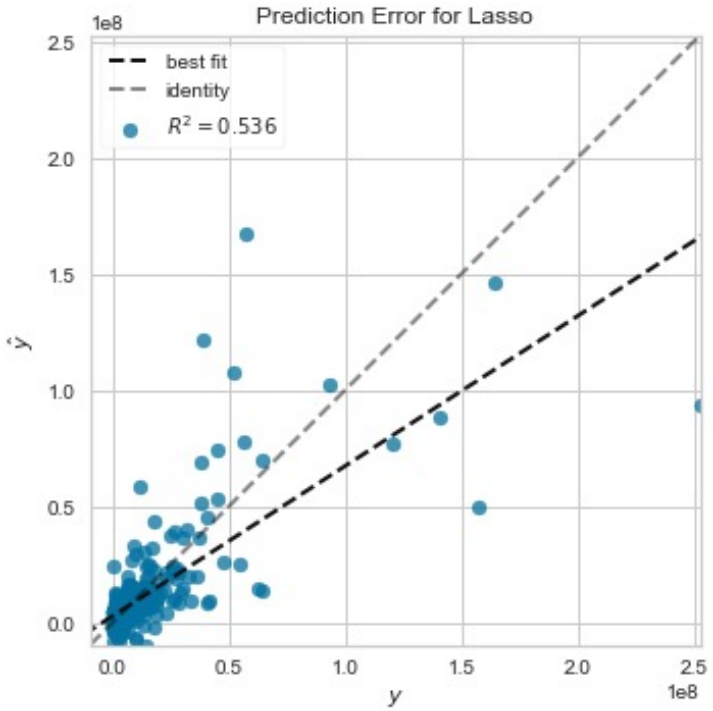
```
Entrée [30]: grid.best_params_

Out[30]: {'alpha': 20.0}
```

- On teste ensuite la prédiction sur le jeu de test : la RMSE sur le jeu de test est $1.31e^{+07}$

(RMSE sur jeu d'entraînement = $1.31e^{+07}$)





4. MODÈLE FINAL SELECTIONNÉ : ÉMISSIONS DE CO2 SANS ENERGYSTARSCORE

4.2. MODÈLE GRADIENT BOOSTING REGRESSOR OPTIMISÉ

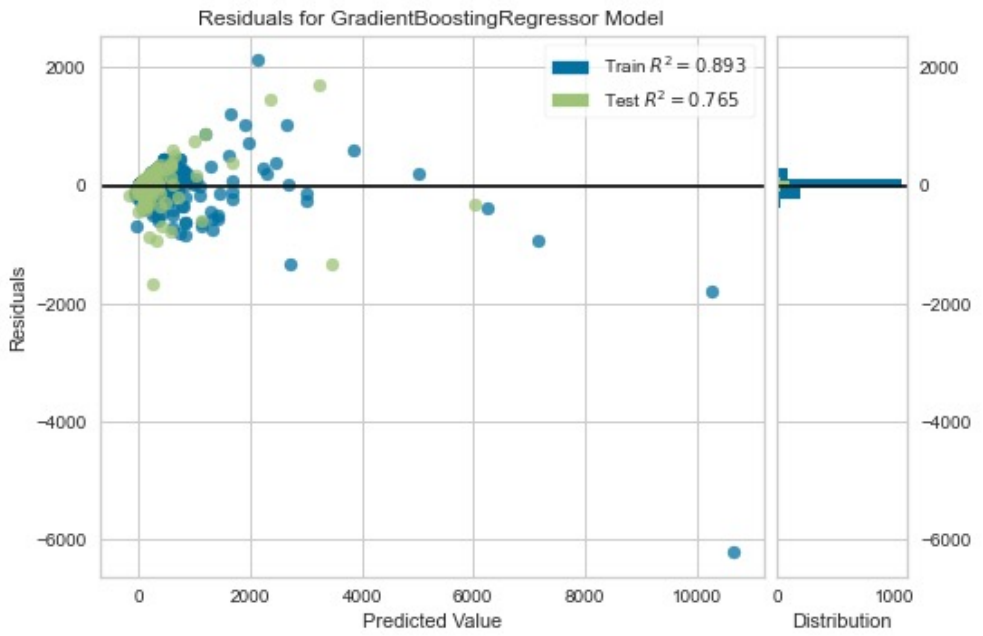
- On cherche les hyperparamètres du Gradient Boosting Regressor pour affiner la performance du modèle :

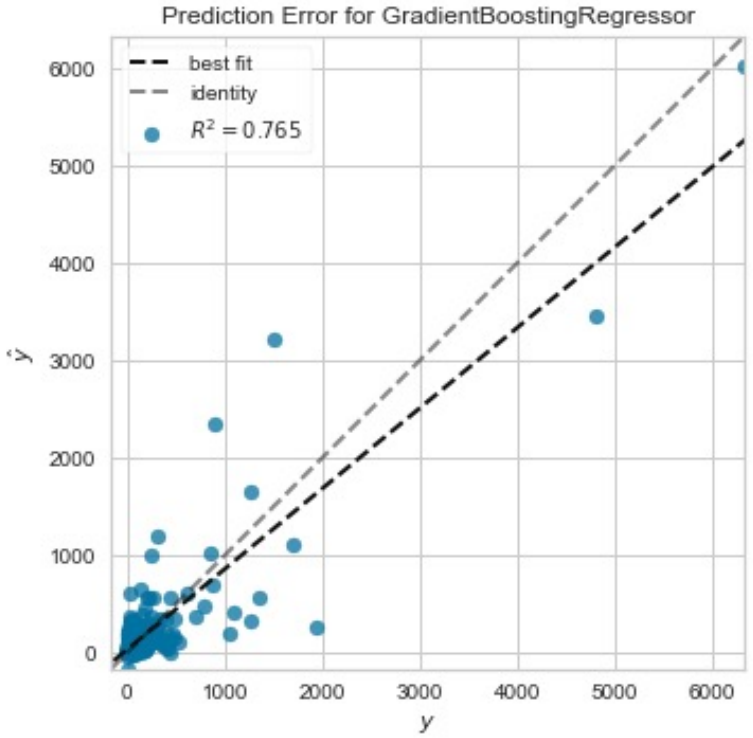
```
Entrée [61]: grid_search.best_params_

Out[61]: {'max_features': 1,
          'min_samples_leaf': 1,
          'min_samples_split': 10,
          'n_estimators': 1000}
```

- On teste ensuite la prédiction sur le jeu de test : la RMSE sur le jeu de test est 229.2

(RMSE sur jeu d'entraînement = 199.4)





4. MODÈLE FINAL SELECTIONNÉ : ÉMISSIONS DE CO2 AVEC ENERGYSTARSCORE

4.3. MODÈLE GRADIENT BOOSTING REGRESSOR OPTIMISÉ

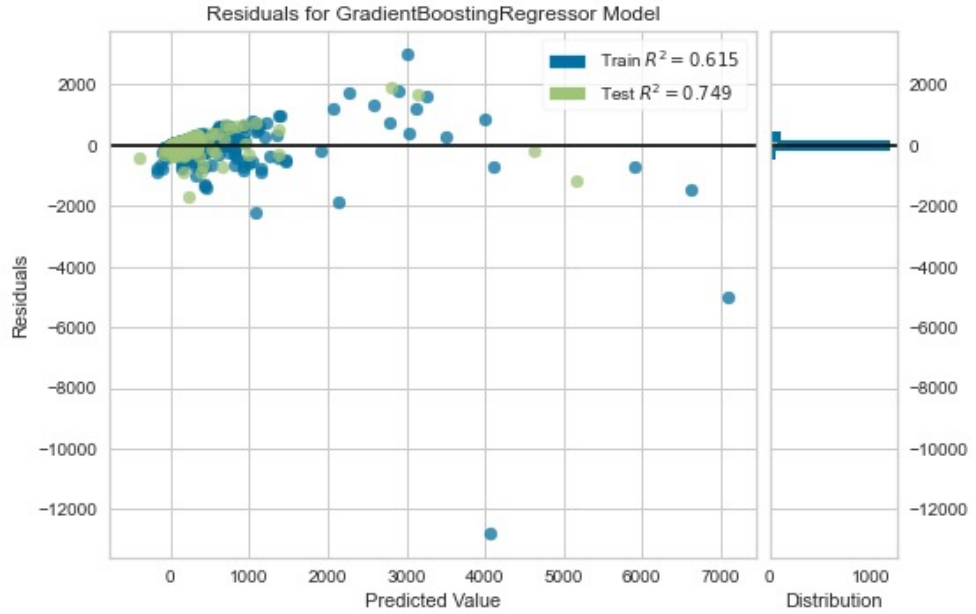
- On cherche les hyperparamètres du Gradient Boosting Regressor pour affiner la performance du modèle.

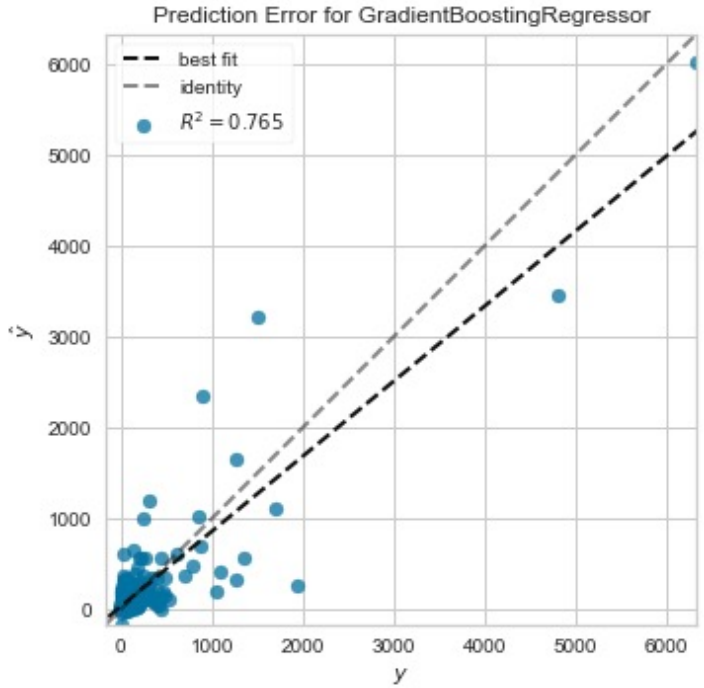
```
Entrée [84]: grid_search.best_params_

Out[84]: {'max_features': 1,
          'min_samples_leaf': 5,
          'min_samples_split': 50,
          'n_estimators': 1000}
```

- On teste ensuite la prédiction sur le jeu de test : la RMSE sur le jeu de test est 237.

(RMSE sur jeu d'entraînement = 198.7)





4. MODÈLES FINAUX SELECTIONNÉS :RÉCAPITULATIF

4.4. CONCLUSION

Grandeur à prédire	**Energie Consommée** *SiteEnergyUse(kBtu)*		**Emissions de CO2 SANS ENERGYSTAR Score** *TotalGHGEmissions*		**Emissions de CO2 SANS ENERGYSTAR Score** *TotalGHGEmissions*	
Modèle le plus performant	LASSO (modèle linéaire)		GradientBoostingRegressor (modèle ensembliste)		GradientBoostingRegressor (modèle ensembliste)	
Jeu de données	entraînement	test	entraînement	test	entraînement	test
Hyperpara-mètres / **Critère**	défault	alpha/lambda = 20	défault	max_features : 1 min_samples_leaf : 1 min_samples_split : 10 n_estimators : 1000	Défault	max_features : 1 min_samples_leaf : 5 min_samples_split : 50 n_estimators : 1000
RMSE	$1.31e^{+07}$	$1.31e^{+07}$	199.4	229.2	198.7	237
R^2	0.535	0.536	0.822	0.765	0.824	0.749

- Les valeurs des RMSE sur les jeux de tests sont dans les mêmes ordre de grandeur que sur les jeux d'entraînement donc les modèles n'ont pas réalisé de sur-apprentissage.
- Il n'y pas d'impact de l'ENERGYSTAR Score sur la prédiction des émissions de CO2.